Dépendances pour Pratique
Introduction
Présentation IBS/IBD
Aperçu des données
Objectifs pratiques
1 Importation de données
1.1 CQ initial
1.2 Exploration des hets
2 Intuition
2.1 Visualiser la recombinaison
3 Calculs
3.1 Identité par État
3.2 Identité par descendance
4 IBD appliqué
5 IBD appliqué
5.1 Intensité de transmission
5.2 Isolement par distance
5.3 Paires apparentées élevées

Calculating IBS & IBD

Code

Nick Brazeau, Sophie Berube, Izzy Routledge, & Abebe Fola

August 05, 2022

Dépendances pour Pratique

Veuillez copier et coller le morceau de code ci-dessous dans son intégralité sur votre console pour télécharger les bibliothèques de packages R nécessaires à cette pratique. Si vous rencontrez des difficultés pour installer l'un des packages R, veuillez demander à un instructeur un lecteur flash préchargé.

Hide

```
deps <- c("tidyverse", "vcfR", "MIPanalyzer", "hmmibdr", "sf", "tidygraph", "ggraph")
deps <- !sapply(deps, function(x){x %in% installed.packages()[,1]} )
if(any(deps)) {
  if(deps["hmmibdr"]) {
    if (!"remotes" %in% installed.packages()[,1]){
      install.packages("remotes")
    }
    remotes::install_github("OJWatson/hmmibdr")
    deps <- deps[names(deps) != "hmmibdr"]
  } else if(deps["MIPanalyzer"]) {
    if (!"remotes" %in% installed.packages()[,1]){
      install.packages("remotes")
    }
    remotes::install_github("mrc-ide/MIPanalyzer")
    deps <- deps[names(deps) != "MIPanalyzer"]
  } else {
    install.packages(names(deps)[deps])
  }
}
```

Veuillez maintenant charger toutes ces bibliothèques dans cette session en utilisant le morceau de code ci-dessous. Veuillez le copier-coller dans son intégralité.

Hide

```
library(tidyverse)
library(vcfR)
library(hmmibdr)
library(sf)
library(tidygraph)
library(ggraph)
library(cowplot)
```

Enfin, veuillez sourcer (*c'est-à-dire* charger) le fichier appelé `utils.r` qui est stocké sous le répertoire `IBD/R`. Si vous utilisez l'environnement `IBD.Rproj`, vous pouvez simplement exécuter `source("R/utils.R")` comme ci-dessous. Ou si vous exécutez à partir d'un environnement ou d'un répertoire de travail différent (`getwd()`), utilisez la fonction `file.choose` pour vous aider à localiser le fichier.

Hide

```
source("R/utils.R")
```

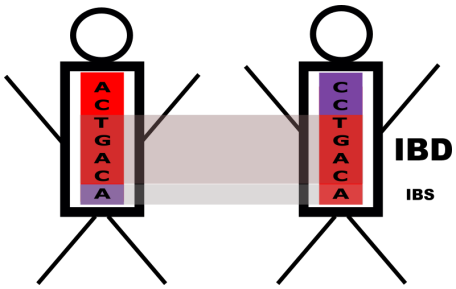
Introduction

Aperçu de la relation {.unnumbered}

Au cours des sessions précédentes, vous avez exploré comment estimer les relations en fonction des niveaux de diversité, de structure et de connectivité de la population. Dans cette pratique, nous nous concentrerons sur les mesures inter-parasites, ou par paires, de la parenté génétique. Il s'agit de mesures de parenté plus axées sur "l'individu" par rapport à la population, ou dème, des mesures de parenté axées sur la population.
Il existe deux façons principales d'estimer la parenté entre les individus: l'identité par descendance (IBD) ou le temps jusqu'à l'ancêtre commun le plus récent (TMRCA) (voir Speed & Balding 2015 (https://pubmed.ncbi.nlm.nih.gov/%2025404112/) pour une discussion plus approfondie). Nous nous concentrerons sur l'IBD, qui tire parti de la recombinaison pour détecter l'héritage commun récent de matériel génétique ou la parenté génétique.

Présentation IBS/IBD

Comme décrit dans la conférence, nous pouvons considérer la parenté par paires en déterminant si la séquence génétique à une position donnée dans le génome, un loci, entre deux parasites est identique (par exemple, le même allèle). Nous pouvons soit simplement mesurer le nombre de sites avec des allèles identiques entre deux individus, appelés identité par état (IBS), soit nous pouvons utiliser des modèles statistiques pour déterminer si des allèles identiques et des "blocs" du génome étaient susceptibles d'être hérités d'un ancêtre commun, appelé identité par descendance (IBD). Il est important de noter les différences entre IBS et IBD: IBS fait référence à "l'état" ou à la réalisation de l'allèle à un loci et peut être le résultat de nombreux facteurs autres que l'ascendance (par exemple, la sélection, la dérive, la structure, le hasard). En revanche, les MII se réfèrent uniquement au matériel génétique hérité, qui suit des schémas spécifiques pertinents pour la santé publique que nous explorerons ci-dessous. Les différences entre IBS et IBD sont visualisées dans le schéma ci-dessous, où les sites peuvent être IBS mais pas IBD, tandis que tous les sites qui sont IBD sont IBS.

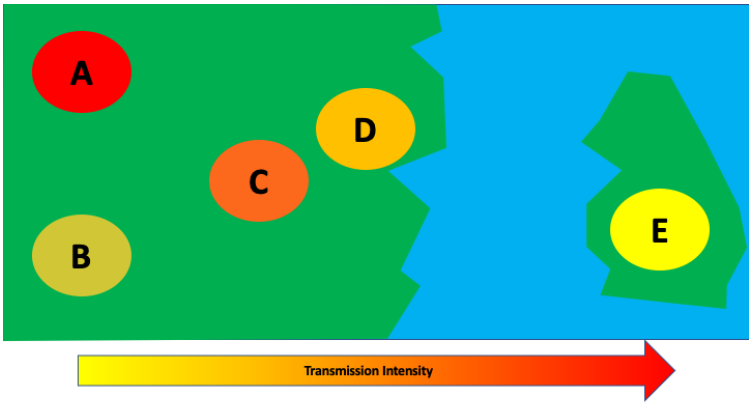


Aperçu des données

Ces données ont été simulées à l'aide d'un modèle simple de paludisme qui prend en compte les relations spatiales, la complexité de l'infection (COI) et la taille de la population sous un modèle de Wright-Fisher: PolySimIBD (https://github.com/nickbrazeau/polySimIBD) - initialement publié dans [Verity/Aydemir/Brazeau et al. 2020]

(https://pubmed.ncbi.nlm.nih.gov/32355199/ (https://pubmed.ncbi.nlm.nih.gov/32355199/)). *Remarque*, les données simulées ont un seul chromosome qui ne fait que 1 000 paires de bases (beaucoup plus petit que les chromosomes du paludisme) et, par conséquent, a un taux de recombinaison beaucoup plus faible (pertinent pour les paramètres d'entrée `hmmIBD`.)
Pour ce TP, nous avons simulé cinq populations, ou dèmes : A-E. Les dèmes ont été simulés pour avoir le même nombre de personnes, ou d'hôtes ; cependant, nous avons sélectionné au hasard cinq individus de chaque dème pour subir une surveillance moléculaire et être "séquencés" (*c'est-à-dire* dans la vraie vie, vous pouvez sélectionner cinq participants d'un village pour donner du sang total contre la collecte de sang de chaque individu dans le village). Les dèmes ont une intensité variable de transmission, ou incidence, comme indiqué par le spectre jaune-rouge dans la légende schématique. De plus, nous nous attendons à ce que les moustiques migrent entre les dèmes en fonction de la distance qui les sépare les uns des autres (*i.e.* isolement par la distance). Enfin, nous supposons que l'océan entre les dèmes A-D et E est une barrière à toute transmission du "continent" vers l'"île" (_c'est-à-dire que nous nous attendons à ce que les parasites du continent soient distincts des parasites insulaires). En résumé, les données consistent en :

- Cinq dèmes (4 continentales, 1 île)
- Les emplacements sont connus pour chaque dème
- Cinq échantillons aléatoires de chaque dème
- Incidence estimée par COI moyen par dème*
  **Comme discuté hier, le COI est une mauvaise mesure de l'incidence dans le monde réel. Pour notre modèle, il est assez corrélé (en attente) et sera utilisé comme proxy ici.*



Objectifs pratiques

À la fin de cet exercice pratique, vous devriez être en mesure d'effectuer les tâches ci-dessous et de comprendre les concepts suivants: + Calculer l'IBS + Calculer l'IBD + Comprendre les forces et les faiblesses d'IBS + Comprendre les forces et les faiblesses des MII

---

1 Importation de données

Ici, nous utiliserons le [package] `vcfR` (https://knausb.github.io/vcfR_documentation/ (https://knausb.github.io/vcfR_documentation/)) pour lire dans notre fichier d'appel de variantes (VCF) qui contient cinquante SNP bialléliques pour nos vingt-cinq individus à travers nos cinq populations (A, B, C, D, E ; *individus nommés* A1-5, B1-5, ...).

En utilisant les compétences que vous avez acquises hier, lisez dans le VCF à partir de notre dossier de données que vous avez téléchargé depuis Github: `data/simulated_ibd.vcf.gz`. Si vous rencontrez des difficultés pour localiser le fichier sur votre ordinateur, veuillez saisir `file.choose` dans votre console pour une option interactive permettant de localiser le fichier. Nommez l'objet (le fichier vcf que vous lisez) "vcf" pour plus de simplicité.

Hide

```
# participants DO NOT need to include the quiet bit
vcf <- quiet(vcfR::read.vcfR("data/simulated_ibd.vcf.gz"))
```

1.1 CQ initial

Comme c'est toujours la bonne pratique, nous pouvons vérifier si notre VCF contient des données manquantes et confirmer qu'il contient le nombre de SNP et d'échantillons que nous attendons. Explorez les détails et le fonctionnement interne de notre nouveau VCF pour savoir s'il est digne de confiance (*c'est-à-dire* les données semblent-elles raisonnables? Y a-t-il quelque chose d'étrange dans la colonne INFO?).

Question de codage 1: *Utilisez la fonction "extract.gt" du package vcfR pour extraire la profondeur d'allèle du premier enregistrement (allèle référent) pour votre nouveau VCF. Tracez maintenant cette profondeur d'allèle en tant qu'exploration initiale des données à l'aide de la fonction "heatmap.bp" du package "vcfR" pour générer la figure ci-dessous.*

Click For Answer

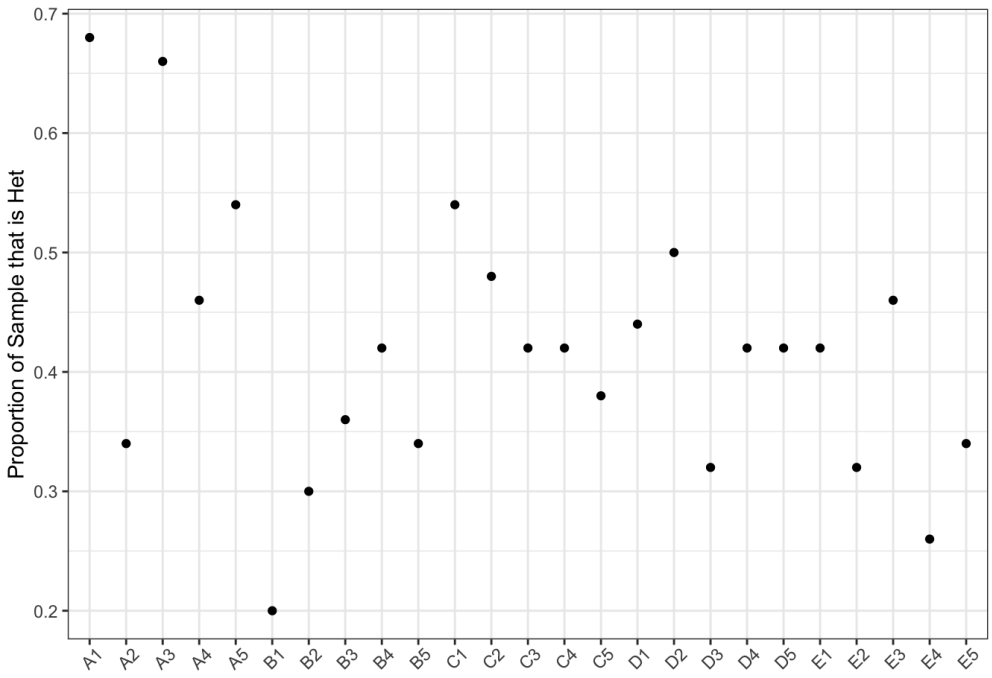
Question conceptuelle 1: *Le tracé ci-dessus est-il cohérent avec un VCF de haute qualité? Quelles sont les fonctionnalités douteuses?*

Click For Answer

1.2 Exploration des hets

Ici, nous allons explorer le nombre d'appels de génotype hétérozygote que nous avons par échantillon dans notre VCF. N'oubliez pas que les appels de génotype hétérozygote nous donnent une approximation grossière du COI.

Question de codage 2: Encore une fois, utilisez la fonction `extract.gt` du package vcfR et extrayez les appels de génotype (`element = "GT"`) pour tous les échantillons. Ensuite, placez les données dans un format "long" en utilisant `pivot` plus longtemps. Une fois que vous avez des données "très longues", calculez l'hétérozygotie moyenne par échantillon à l'aide de la fonction "résumer" et tracez les données dans un nuage de points, comme ci-dessous.



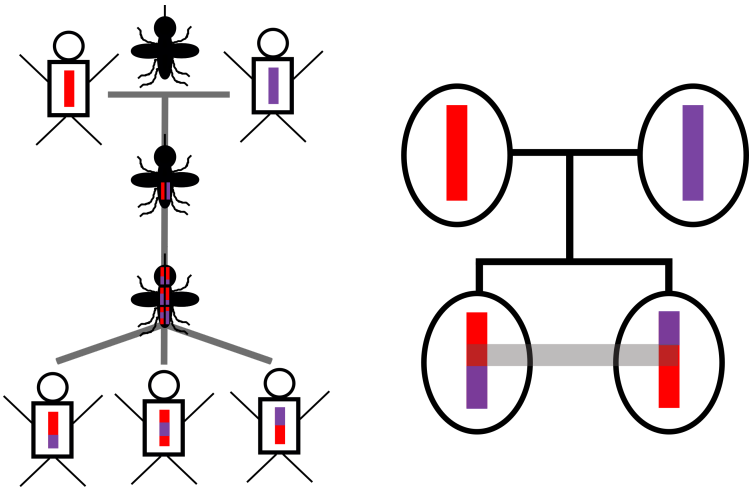
Click For Answer

Question conceptuelle 2: *Reconnaissez-vous un schéma dans les appels hétérozygotes en ce qui concerne les dèmes? Quel est le rapport avec la leçon sur la polyclonalité d'hier?*

Click For Answer

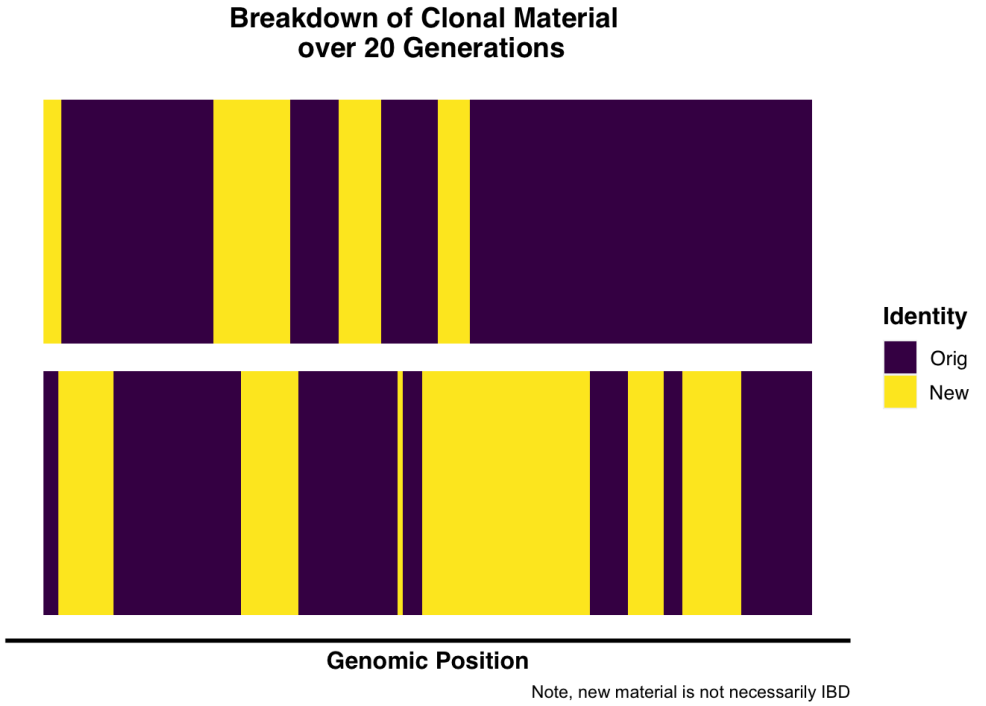
2 Intuition

La recombinaison produit un signal puissant pour détecter la parenté génétique entre les individus. Dans le paludisme, la recombinaison se produit dans l'intestin moyen du moustique. Pour rappel, le schéma ci-dessous montre un seul moustique qui a piqué deux individus différents infectés par une même souche du parasite du paludisme, entraînant une surinfection. Les deux souches (rouge vs violet) subissent ensuite une recombinaison dans l'intestin moyen du moustique et produisent des descendants qui sont frères et devraient partager en moyenne la moitié de leurs génomes (*ils sont séparés par 1 génération*). L'arbre généalogique de droite met en évidence deux des frères et sœurs du parasite et comporte une barre grise indiquant la quantité de leur génome héritée du même parent au même emplacement génomique.



2.1 Visualiser la recombinaison

En conséquence, une façon de conceptualiser le génome est de le faire sous forme de blocs, ou de segments, qui sont reliés entre eux, où chaque bloc a sa propre histoire ancestrale unique. Ainsi, la recombinaison peut être considérée comme le processus de "casse-bâton" qui brise les segments du génome/blocs IBD: à mesure que le temps augmente, ou que le nombre de générations augmente, les blocs génomiques partagés entre les individus deviennent plus petits. Par exemple, vous pouvez imaginer que vous avez deux sticks identiques (*i.e.* jumeaux ou infections clonales). À chaque nouvelle génération, vous devez casser votre bâton au hasard à un moment donné de sa longueur. Vous prenez ensuite vos morceaux de bâton et lancez une pièce pour déterminer si elle restera la même (« originale ») ou héritera du matériel génétique (« nouvelle »). Utilisez la fonction `view_recombo` (chargée à partir de `R/utils.R`) et explorez la relation entre le nombre de générations, ou le temps, et la longueur des blocs de recombinaison qui sont IBD entre les clones. La figure ci-dessous montre les résultats pour 20 générations.



Par exemple, vous pouvez exécuter la fonction avec les durées générationnelles suivantes

Hide

```
view_recombo(generations_apart = 0)
view_recombo(generations_apart = 1)
view_recombo(generations_apart = 2)
view_recombo(generations_apart = 3)
view_recombo(generations_apart = 5)
view_recombo(generations_apart = 10)
view_recombo(generations_apart = 20)
```

Question conceptuelle 3: *Comment la quantité de matériel génétique "clonal" change-t-elle, ou la quantité de génome identique, change-t-elle entre les échantillons appariés à mesure que nous augmentons le nombre de générations qui les séparent?*

Click For Answer

3 Calculs

3.0.1 Manipulation interne des données

Ici, nous allons convertir notre objet *vcf* que nous avons lu avec le package `vcfR` en un objet *mip* afin d'utiliser ultérieurement le `MIPAnalyzer` Package (https://github.com/mrc-ide/MIPanalyzer%20/tree/master/R) et l'estimateur du maximum de vraisemblance de l'IBD de [Verity et al. 2020] (https://pubmed.ncbi.nlm.nih.gov/32355199/ (https://pubmed.ncbi.nlm.nih.gov/32355199/)). Ceci est purement pratique et ne modifie pas les données sous-jacentes dans le vcf.Veuillez copier et coller le code ci-dessous dans votre console.

Hide

```
mipvcf <- MIPanalyzer::vcf2mipanalyzer_biallelic(vcfR = vcf)
```

3.1 Identité par État

L'identité par état (IBS) est la proportion de loci identiques divisé par tous les loci mesurés dans le génome entre deux parasites: parasite$_a$ et parasite$_b$, tel que:

$$IBS = \frac{\text{Lieux partagés}_{ab}}{\text{Nombre de lieux}}$$

Ici, nous allons effectuer des calculs IBS en utilisant notre VCF comme entrée.

Question de codage 3: Choisissez deux échantillons de votre VCF et écrivez votre propre fonction pour calculer l'IBS entre la paire. Je recommande de commencer avec l'objet `vcfR` (par opposition à l'objet `mipanalyzer_biallelic` nouvellement créé) et d'utiliser la fonction `extract.gt` du package vcfR pour extraire les appels de génotype (`element = "GT"`) comme ci-dessus. Ensuite, subdivisez les données en deux échantillons et comparez les lieux.

Click For Answer

3.1.1 Exécuter le package IBS

Ensuite, nous utiliserons le calculateur IBS du package `MIPAnalyzer` R pour calculer l'IBS par paires pour toutes les combinaisons d'échantillons dans nos données simulées. Veuillez copier et coller le code ci-dessous dans votre console pour faciliter l'exécution de l'algorithme.

Hide

```
# get IBS
ibs <- MIPanalyzer::get_IBS_distance(x = mipvcf,
                                     ignore_het = FALSE,
                                     report_progress = FALSE)
```

3.1.1.1 Résultats nets

La fonction renvoie les données dans un format "large", ou une matrice de distance (https://en.wikipedia.org/wiki/Distance_matrix). Bien qu'il s'agisse d'un format parfaitement acceptable, il n'est pas considéré comme « bien rangé ».

Question de codage 4:: agencez les données au format "long" à l'aide de la fonction `broom::tidy` . Renommez les colonnes pour les données longues en `c("p1", "p2", "ibsdist")` . Remarque, assurez-vous de conserver les noms d'échantillons de `colnames(vcf@gt)[2:ncol(vcf@gt)]` avant de ranger !

Click For Answer

Il est toujours bon d'explorer nos résultats. Faisons une boîte à moustaches pour explorer la distribution de nos résultats IBS calculés.
Question de codage 5:: Terminé le code ci-dessous pour créer une boîte à moustaches avec des valeurs IBS sur l'axe y.

```
ibs_long %>%
  ggplot() +
  geom_boxplot(***) +
  ylab("IBS Values") +
  theme_bw() +
  theme(axis.text.x = element_blank(),
        axis.ticks.x = element_blank())
```

Click For Answer

Question conceptuelle 4: *Décrivez la distribution. Est-ce ce que vous attendiez?*

Click For Answer

3.2 Identité par descendance

L'identité par descendance (IBD) est plus compliquée à calculer que l'IBS et nécessite une modélisation statistique afin de tenir compte des fréquences de recombinaison et d'allèles de la population. Le modèle statistique le plus couramment utilisé est appelé modèle de Markov caché (voir la section * Lectures complémentaires * pour les sources recommandées). N'oubliez pas qu'un site peut être IBS mais pas nécessairement IBD comme décrit dans la section [Aperçu IBS/IBD].
Ici, nous utiliserons deux algorithmes différents pour calculer l'IBD : (1) un estimateur de consanguinité MLE ; (2) un modèle de Markov caché.

3.2.0.1 MLE MII

Ce premier algorithme utilise un estimateur du maximum de vraisemblance des MII (voir Verity et al. 2020 (https://pubmed.ncbi.nlm.nih.gov/33057671/) pour une description mathématique). Le modèle suppose que les loci sont indépendants, ce qui permet de calculer essentiellement l'IBD comme la quantité de consanguinité, ou l'écart par rapport aux fréquences alléliques attendues, entre les fréquences alléliques par site. Le modèle est une variante de l'équation de Hardy-Weinberg (https://en.wikipedia.org/wiki/Hardy%E2%80%93Weinberg_principle) généralisée pour l'accouplement non aléatoire (voir [Gillespie, Population Genetics: A Concise Guide] (https://public.wsu.edu/~gomulki/mathgen/materials/gillespie_book.pdf (https://public.wsu.edu/~gomulki/mathgen/materials/gillespie_book.pdf)) 1ère édition, pg 86, pour en savoir plus).
Comme ci-dessus, nous utiliserons le package R "MIPAnalyzer" pour calculer l'IBD par paires pour toutes les combinaisons d'échantillons dans nos données simulées. Pour plus de commodité, copiez et collez le code ci-dessous dans votre console pour acquérir ces résultats.

Hide

```
#.....................
# MLE IBD
#.....................
ibd <- MIPanalyzer::inbreeding_mle(x = mipvcf,
                                   f = seq(0.01, 0.99, 0.01),
                                   ignore_het = FALSE,
                                   report_progress = FALSE)
```

Question de codage 6:: Rangeons/convertissons à nouveau nos données au format "long" en utilisant la fonction `broom::tidy` . Assurez-vous de définir votre diagonale sur 1 (auto-comparaisons), de conserver les noms des échantillons et de nommer vos colonnes `c("p1", "p2", "malecotf")` .

Click For Answer

Après avoir généré nos nouveaux résultats, nous les explorons toujours!
Question de codage 7: À l'aide du code `geom_boxplot` ci-dessus, créez une boîte à moustaches avec les valeurs IBD sur l'axe des ordonnées.

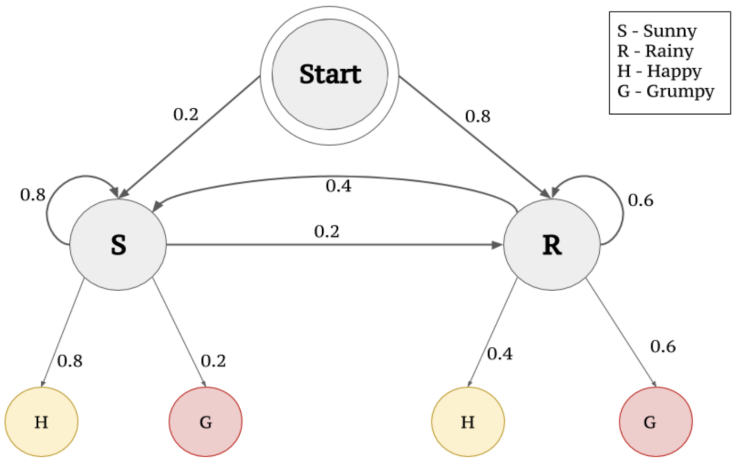
Click For Answer

Question conceptuelle 5: *Décrivez la distribution MLE. Est-ce ce que vous attendiez?*

Click For Answer

3.2.0.2 IBD de HMM

Ensuite, nous utiliserons un modèle de Markov caché (HMM) du R-package `hmmIBDR` , qui est un wrapper pour l'algorithme original `hmmIBD` (pour une description mathématique complète de `hmmIBD` voir Schaffner & Taylor 2018 (https://pubmed.ncbi.nlm.nih.gov/29764422/)) pour calculer les estimations de l'IBD. Comme décrit dans la conférence, les HMM peuvent être considérés comme un modèle statistique qui utilise des "observations" pour déduire l'état "caché" d'un processus, où le processus est connu ou peut être décrit. Dans notre cas, nos observations sont des appels de génotype entre individus (sites identiques versus sites discordants) et notre état caché est de savoir si oui ou non ce site est une région d'héritage commun entre individus (IBD) ou non. Le processus sous-jacent est la recombinaison, qui, nous le savons, rompt les blocs génomiques et peut être décrite sur la base du taux de recombinaison connu dans le paludisme. Ce site (https://towardsdatascience.com/markov-and-hidden-markov-model-3eec42298d75) fournit un exemple non génomique utile: imaginez que vous appelez chaque jour un ami qui habite loin et que cet ami a deux «état» d'esprit: grincheux ou heureux. Vous savez que l'humeur de votre ami dépend de la météo, mais nous ne connaissons pas la météo dans cet endroit lointain (elle nous est "cachée"). Cependant, nous pouvons déduire le temps qu'il faisait ce jour-là en fonction du fait que notre ami était heureux ou grincheux lorsque nous lui avons parlé. De plus, si nous enregistrons l'humeur de notre ami pendant une semaine, nous pourrions utiliser un modèle pour déduire le modèle météorologique sur toute cette semaine en fonction de la séquence observée de "l'état d'esprit" de notre ami, comme indiqué dans le schéma ci-dessous. Par exemple, si notre ami était G, G, G, H, H, H, H, nous pouvons soupçonner qu'il a plu pendant les trois premiers jours de la semaine, mais ensuite il y a eu un changement d'état et il a fait beau les quatre derniers jours de la semaine.



3.2.0.2.1 Formatage pour `hmmIBD`

Avant d'exécuter le [programme] `hmmIBD` (https://github.com/glipsnort/hmmIBD (https://github.com/glipsnort/hmmIBD)), nous allons devoir procéder à quelques manipulations de données.

Question de codage 8, partie 1: Créez une matrice de fréquence allélique alternative (WSAF) intra-échantillon. Tout d'abord, commencez par extraire l'allèle alternatif, ou le nombre de lectures alternatives, à un locus particulier par échantillon à l'aide de la fonction `vcfR::extract.gt(vcf, element = "AD")` : (le numérateur). Ensuite, calculez la profondeur totale des allèles par locus, par échantillon à l'aide de la fonction `vcfR::extract.gt(vcf, element = "DP")` : (le dénominateur). Maintenant, trouvez le WSAF en divisant le numérateur par le dénominateur et nommez cette nouvelle matrice `wsaf` .

Click For Answer

Coding Question 9, part 2: Afin de convertir cette trame de données en une trame de données conforme à " `hmmIBD` , ajoutez d'abord le chromosome et la position à votre matrice wsaf (en tant que première et deuxième colonnes respectivement) et le nom il `gtmat` . Ensuite, rangez votre matrice"large" au format "long" en excluant le chromosome et la position (*c'est-à-dire* nous voulons une base de données avec chromosome, position, échantillon, wsaf comme colonnes) avec la fonction: `tidyr :: pivot_longer(., cols = !c("chrom", "pos"))` . Nous allons maintenant contraindre nos calculs WSAF en appels de génotypes homozygotes en utilisant l'arrondi: `gthmm = round(value, 0)` . De plus, assurez-vous de supprimer la colonne `value` (maintenant redondante) et pour `hmmIBD` de remplacer les noms de chromosomes par des entiers: (`chrom = 1)` . Par la suite, nous allons maintenant rendre nos données "larges" à partir de longues et écrivez cette nouvelle table sur votre disque local pour `hmmIBD` en utilisant les fonctions:
`tidyr::pivot_wider(data = ., names_from = "name", values_from = "gthmm")` et
`readr::write_tsv(x = gtmat, fichier = " data/gt_matrix_for_hmmIBD.txt", col_names = T)` , respectivement.

Click For Answer

Question de codage 10, partie 3 : Enfin, à l'aide de votre matrice `wsaf` , calculez les fréquences alléliques au niveau de la population pour chaque locus à l'aide de la fonction suivante: `altafvec <- rowMeans(wsaf, na.rm = T)` . N'oubliez pas qu'il s'agit de la fréquence d'allèle alternative, et parce que nous avons affaire à des sites bialléliques, nous pouvons calculer la fréquence d'allèle référent en tant que 1-alt et placer le tout dans une base de données. La trame de données a besoin d'une colonne de chromosomes (sous la forme d'un nombre/entier [`chrom = 1`]), d'une colonne pour la position de la paire de bases du locus, d'une colonne de fréquence d'allèle référent et d'une colonne de fréquence d'allèle alternative. Écrivez cette table sur votre disque local avec le chemin suivant: `data/af_matrix_for_hmmIBD.txt` .

Click For Answer

Explorez la documentation hmmIBD (https://github.com/glipsnort/hmmIBD) pour une justification de ce munging que nous avons fait ci-dessus et une clarification supplémentaire de la mise en garde ci-dessous.

Avertissement*: L'algorithme `hmmIBD` suppose que tous les échantillons sont monoclonaux. De ce fait, il n'accepte pas nativement les appels hétérozygotes. En tant qu'utilisateur, nous avons deux options : (1) définir ces appels hétérozygotes comme manquants (une approche conservatrice), ou (2) faire une hypothèse forte sur ce que signifie un appel hétérozygote (une approche agressive). Ici, j'ai forcé les échantillons à être monoclonaux en estimant leurs fréquences alléliques dans l'échantillon et en arrondissant à la variante la plus proche.*

3.2.0.2.2 Exécution hmmIBD

Nous allons maintenant exécuter le programme `hmmIBD` pour calculer l'IBD par paires. Pour plus de commodité, copiez et collez le code ci-dessous dans votre console pour acquérir ces résultats.

Hide

```
#.....................
# hmmIBD run
#.....................
# NB, participants do not need to use the "quiet" function
tf <- tempfile(pattern = "output_simdat")
out <- quiet(hmmibdr::hmm_ibd(input_file = "data/gt_matrix_for_hmmIBD.txt",
                              allele_freqs =  "data/af_matrix_for_hmmIBD.txt",
                              rec_rate = 1e-2, # note the small recombo rate relative to what wou
        ld be expected in malaria
                              output_file = tf))

# hmmIBD tidy
ibd_hmm_long <- tibble::tibble(
  p1 = out$fract$sample1,
  p2 = out$fract$sample2,
  hmm = out$fract$fract_sites_IBD)
```

Encore une fois, nous explorons toujours nos résultats!
Question de codage 11: utilisez votre code `geom_boxplot` pour créer une boîte à moustaches avec les valeurs IBD HMM sur l'axe des ordonnées.

Click For Answer

Question conceptuelle 6: *Décrivez la distribution HMM. Est-ce ce que vous attendiez?*

Click For Answer

3.2.0.3 Contraste MLE vs HMM

Comme décrit ci-dessus, les calculs IBD du modèle MLE supposent des lieux indépendants, tandis que le modèle HMM exploite la recombinaison pour détecter la parenté. Cependant, "hmmIBD" nécessite des infections monoclonales et nous avons dû "contraindre" nos données pour qu'elles correspondent à cette hypothèse. Ici, nous allons explorer comment les résultats MLE IBD et HMM IBD diffèrent.

Question de codage 12: Utilisez la couche ggplot `geom_point` pour créer un tracé qui met en contraste les deux calculs différents. Vous devrez joindre les données par échantillon à l'aide de la fonction suivante:
`left_join(ibd_mle_long, ibd_hmm_long, by = c("p1", "p2"))` en terminant le morceau de code ci-dessous.

```
dplyr::left_join(***) %>%
  ggplot() +
  geom_point(***) +
  theme_bw() +
  labs(x = "IBD from HMM", y = "IBD from MLE")
```

Click For Answer

Question conceptuelle 7: *Pourquoi ces deux résultats seraient-ils différents? Existe-t-il des hypothèses différentes dans les modèles?*

Click For Answer

3.2.0.4 Contraste IBS vs HMM-IBD

Comparez IBS et les résultats `hmmIBD` dans un nuage de points. ***Question de codage 12***: À l'aide de la couche ggplot "geom_point", créez un graphique qui met en contraste les estimations IBS et les estimations MLE IBD. Ajoutez une ligne de régression pour explorer la corrélation à l'aide de la couche `geom_smooth(aes(x = malecotf, y = ibsdist), method = "lm")`. Notez que vous devrez joindre les données par exemple à l'aide de la fonction suivante: `left_join(ibs_long, ibd_mle_long, by = c("p1", "p2"))` avant d'utiliser ggplot.

Click For Answer

Question conceptuelle 8: *En quoi les estimations de l'IBS et de l'IBD diffèrent-elles? Quelle est la plus petite valeur pour IBS par rapport à IBD? Pourquoi peuvent-ils être différents?*

Click For Answer

3.2.1 Résumé/Enregistrement

À ce stade, vous devriez avoir créé trois objets de résultat R:

```
ibs_long
ibd_mle_long
ibd_hmm_long
```

Si vous rencontrez des problèmes de calcul ou si le temps est un facteur limitant, veuillez consulter le répertoire `results/`, où vous pourrez lire les résultats.

4 IBD appliqué

Dans cette section, nous nous concentrerons sur l'utilité de l'IBD pour des applications réalistes liées aux efforts de contrôle et d'élimination. Pour cette section, nous nous concentrerons sur les estimations MLE-IBD (par rapport aux estimations HMM-IBD).

5 IBD appliqué

Dans cette section, nous nous concentrerons sur l'utilité de l'IBD pour des applications réalistes liées aux efforts de contrôle et d'élimination. Pour cette section, nous nous concentrerons sur les estimations MLE-IBD (par rapport aux estimations HMM-IBD).

Question de codage 13: Utilisez la couche ggplot `geom_histogram` pour créer un graphique qui montre la distribution des estimations MLE IBD. Nous ajouterons un graphique en encart pour explorer la queue de cette distribution. Explorez ce blog (https://meghan.rbind.io/blog/cowplot/) pour apprendre à créer des tracés avec des encarts à l'aide du package "cowplot".

Click For Answer

Question conceptuelle 9: *Comment décririez-vous cette distribution? Vous attendriez-vous à autant d'échantillons non liés? Qu'en est-il de la proportion de paires hautement liées dans la queue de la distribution?*

Click For Answer

5.1 Intensité de transmission

Comme nous l'avons vu lors de l'examen de la recombinaison au début de la pratique, à mesure que le nombre d'événements de recombinaison augmente, nous nous attendons à ce que l'IBD diminue. En conséquence, dans les zones où l'intensité de transmission est plus élevée, modélisée ici par des différences de COI, nous pouvons nous attendre à ce que l'IBD soit moindre.

5.1.1 Intra-Deme IBD vs COI

En regroupant les échantillons selon leurs dèmes d'origine, nous pouvons estimer la quantité de consanguinité au sein d'un dème, ou l'IBD intra-dème. Ici, nous allons lire dans le fichier `data/metadata.RDS` et calculer l'IBD moyen par dème pour obtenir l'IBD intra-dème. Nous allons ensuite tracer l'IBD intra-dème par rapport au COI moyen simulé.

Coding Question 13, part 1: Lisez les métadonnées à l'aide du code suivant: `readRDS("data/metadata.RDS")`. Combinez ensuite nos métadonnées à la trame de données `ibd_mle_long` en utilisant deux appels `left_join` différents. Notez que vous devrez créer deux dataframes "métadonnées" différents avec une colonne nommée "p1" (échantillon 1) et une avec une colonne nommée "p2" (échantillon 2). Renommez les extensions ".x" et ".y" créées par `left_join` en "_p1" et "_p2", respectivement. Après avoir joint ces données ensemble, vous devriez avoir un nouveau dataframe avec des noms de colonnes: `p1, p2, malectof, deme_p1, coimeans_p1, longnum_p1, latnum_p1, deme_p2, coimeans_pe2, longnum_p2, latnum_p2`. Nous allons utiliser cette base de données pour les prochains défis, alors nommez-la `ibd_mle_long_mtdt`.

Click For Answer

Question conceptuelle 10: *Décrivez votre nouveau cadre de données.*

Click For Answer

Code Question 14, partie 2: Calculer l'IBD intra-dème. Maintenant, créez un sous-ensemble de votre dataframe pour ne contenir que les lignes où les comparaisons par paires proviennent du même dème: `dplyr::filter(deme_p1 == deme_p2)`. Regroupez par dèmes (`group_by(deme_p1)`) et utilisez la fonction `summarise` pour calculer la moyenne intra-dème IBD et le COI moyen (remarque, le COI est le même dans un dème). Utilisez ensuite la couche `geom_point` dans ggplot pour comparer les calculs COI et IBD intra-dème.

Click For Answer

Question conceptuelle 11: *Existe-t-il ici une corrélation entre le COI moyen et l'IBD au sein du même?*

Click For Answer

Question conceptuelle 12: *Pourquoi nos attentes concernant la relation entre les MICI intra-dème et l'intensité de la transmission ne correspondent-elles pas à la réalité?*

Click For Answer

5.2 Isolement par distance

Ensuite, nous explorerons le concept d'isolement par la distance, qui est la théorie selon laquelle à mesure que les paires s'éloignent dans l'espace, elles devraient être moins liées (*c'est-à-dire* des paires plus proches sont plus susceptibles de se reproduire). Ce concept est basé sur des relations spatiales, mais peut également être conceptualisé en tant que temps (les paires qui sont séparées par plusieurs générations de temps sont moins susceptibles d'être liées).

Question de codage 15, partie 1: Calculer la distance du grand cercle (https://en.wikipedia.org/wiki/Great-circle_distance). Tout d'abord, nous lirons dans nos distances GC pré-calculées : `readRDS(data/deme_gc_dist.RDS)`. Utilisez un appel `left_join` pour fusionner ces données tout en créant une nouvelle trame de données appelée `ibd_mle_long_mtdt_dist`. Maintenant, nous allons "répartir" nos distances en convertissant cette valeur continue en une forme discrétisée à l'aide de la fonction "couper":

```
cut(x = c(ibd_mle_long_mtdt_dist$distance),
breaks = c(0, 1e-26, seq(40, 120, by = 40), Inf),
right = F,
labels = c("Within", "40km", "80km", "120km", ">120km"))
```

Click For Answer

Question de codage 16, partie 2 : À l'aide de notre nouveau cadre de données, `ibd_mle_long_mtdt_dist`, regroupez par la colonne `distance_cat` créée avec la fonction `cut` ci-dessus et calculez les résumés suivants (`summarise`): moyenne IBD, écart type IBD, erreur standard IBD, IC inférieur à 95% pour IBD et IC supérieur à 95% pour IBD. Vous pouvez consulter la pratique de puissance, partie 1 pour un rappel sur ces statistiques.

Click For Answer

Question de codage 17, partie3 : à l'aide de la couche "geom_pointrange" dans ggplot, tracez l'IBD moyen sur l'axe des y et la distance GC catégorisée sur l'axe des x. Pour l'esthétique `geom_pointrange`, vous devrez spécifier:
`y = meanIBD, ymin = L95CI, ymax = U95CI`.

Click For Answer

Question conceptuelle 13: *Comment décririez-vous cette relation ? Est-ce que ça fait du sens?*

Click For Answer

5.2.1 Analyse de réseau

Nous pouvons également utiliser des réseaux pour déterminer la connectivité de nos échantillons appariés. Les réseaux nous aident à visualiser les connexions auxquelles nous ne nous attendons peut-être pas et constituent un outil utile pour l'analyse exploratoire des données. De plus, nous pouvons utiliser des algorithmes de détection de communauté (https://towardsdatascience.com/community-detection-algorithms-9bd8951e7dae) pour déterminer quels échantillons peuvent se "regrouper" en fonction de leur parenté. Une façon de conceptualiser les algorithmes de détection de communauté est de considérer que les réseaux mesurent la "popularité" et que les échantillons qui sont des "cliques" devraient se regrouper. Alternativement, un échantillon peut être "ami" avec tout le monde et devenir très central ou important pour le réseau. Ces distinctions peuvent être utiles pour déterminer quels dèmes sont particulièrement connectés et peuvent contribuer à la dynamique puits-source (voir Wesolowksi *et al.* 2018 (https://pubmed.ncbi.nlm.nih.gov/30333020/) pour informations complémentaires sur la dynamique puits-source).
Lisez ce [blog] STDHA (http://sthda.com/english/articles/33-social-network-analysis/136-network-analysis-and-manipulation-using-r (http://sthda.com/english/articles/33-social-network-analysis/136-network-analysis-and-manipulation-using-r)) réseaux et sur les algorithmes de détection de communauté et utilisez le Packages R: tidygraph (https://tidygraph.data-imaginist.com/index.html) et ggraph (https://ggraph.data-imaginist.com/) pour essayer de générer la figure ci-dessous.

Question de codage 18: Utilisez la fonction `tidygraph::as_tbl_graph` pour convertir notre `ibd_mle_long` en un `tbl_graph` afin de faciliter les analyses de réseau. Calculez ensuite l'appartenance à la communauté en utilisant la fonction `tidygraph::group_louvain(weights = malecotf))` et tracez le réseau résultant avec `ggraph::ggraph(layout = 'kk')`. Assurez-vous de colorer les nœuds par communauté en utilisant le calque `geom_node_point` et l'esthétique appropriée: `ggraph :: geom_node_point(aes(color = community)`.

Click For Answer

Question de codage 19 : Bien que ce graphique soit intéressant, il peut être difficile à interpréter compte tenu de toutes les relations de faible niveau. Élaguons-les et ne gardons que les connexions avec IBD $\geq$ 0.1. Utilisez la fonction `dplyr::filter` pour effectuer cette tâche.

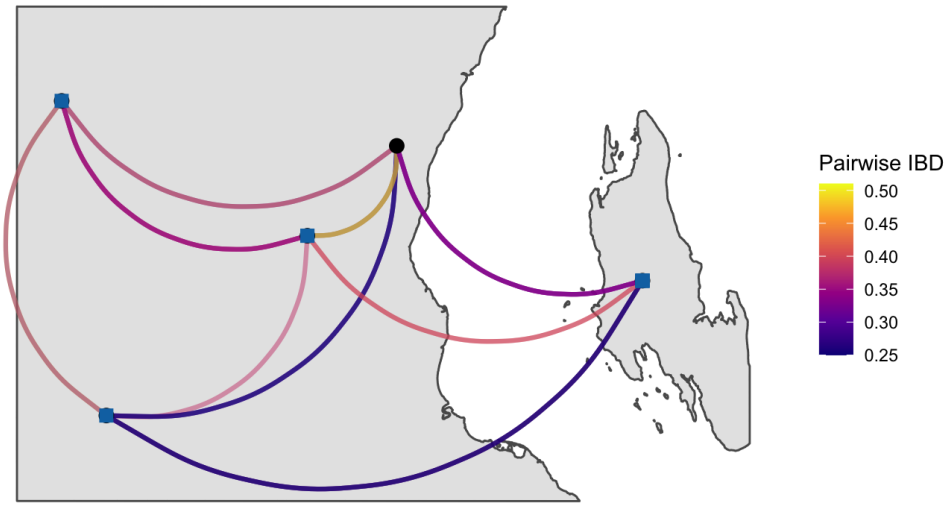
Click For Answer

Question conceptuelle 14: *Que remarquez-vous à propos du réseau ci-dessus? Existe-t-il des cliques isolées ? Existe-t-il des échantillons très "populaires"?*

Click For Answer

5.3 Paires apparentées élevées

Ici, nous allons subdiviser en paires hautement apparentées, définies comme des paires qui sont au moins des frères et sœurs méiotiques (*i.e.* partagent la moitié de leur génome). Nous cartographierons ensuite ces connexions entre les dèmes pour déterminer s'il existe des preuves de partage génétique entre les dèmes. Dans cette section, nous allons pratiquer la superposition de plusieurs dataframes sur un seul tracé pour générer une figure comme ci-dessous.



Locations that contain highly related pairs within
the deme are colored as blue squares;
between deme connections are shown are curved lines

Question de codage 20, partie 1: Ici, nous allons configurer nos données pour créer la carte/le tracé. Lire dans la carte simulée avec la fonction suivante : `readRDS("data/sim_map_sf.RDS")`. Créez ensuite une version condensée des métadonnées qui ne contient qu'une ligne de coordonnées pour chaque dème. Ensuite, créez une base de données contenant des paires hautement liées entre les dèmes en sous-définissant le `ibd_mle_long_mtdt` à au moins des frères et sœurs méiotiques: `dplyr::filter(malecotf >= 0,25)` et en excluant les paires qui proviennent du même dème: `dplyr:: filtre(dème_p1 != dème_p2)`. Enfin, créez une base de données de dèmes contenant des paires hautement liées en sous-définissant `ibd_mle_long_mtdt` au moins sur les frères et sœurs méiotiques: `dplyr::filter(malecotf >= 0,25)` et en incluant uniquement les paires provenant du même dème: `dplyr ::filter(dème_p1 == dème_p2)`. Vous savez que quatre objets sont prêts à être tracés: (1) carte, (2) emplacements de clusters, (3) fortement liés entre les paires de dèmes, (4) fortement liés au sein des paires de dèmes.

Click For Answer

Question de codage 21, partie 2 : Nous allons maintenant tracer ces trames de données. Créez d'abord une base de carte en utilisant la couche `geom_sf`. Ajoutez ensuite une couche pour les emplacements des clusters en utilisant `geom_point(data = mtdtclst, aes(x = longnum, y = latnum)`. Maintenant, nous allons colorier les dèmes qui contiennent des paires hautement liées dans: `geom_point(data = withinpairs, aes( x = longnum_p1, y = latnum_p1), color = "blue"`. Enfin, nous ajouterons une connexion entre les paires à travers différents dèmes à partir de notre cadre de données hautement lié entre les paires: `geom_curve(data = highbtwnpairs, aes(x = longnum_p1, y = latnum_p1, xend = longnum_p2, yend = latnum_p2, color = malecotf)". Votre graphique contient à la fois des données spatiales et de relation dans un graphique facilement communiqué. Notez que vous pouvez modifier l'ordre des couches afin d'avoir des points ou des courbes sur top (rappelez-vous, les couches ggplot suivent l'ordre du code).

Click For Answer

Question conceptuelle 15: *Que peuvent vous dire les paires hautement apparentées? Comment sont-ils liés aux événements de transmission? Pouvez-vous les utiliser pour estimer les corridors génétiques ou les preuves de flux de gènes ? Pensez-vous qu'il y a une transmission active entre le continent et l'île (voir hypothèse d'étude originale)?*

Click For Answer